FTML practical session 5

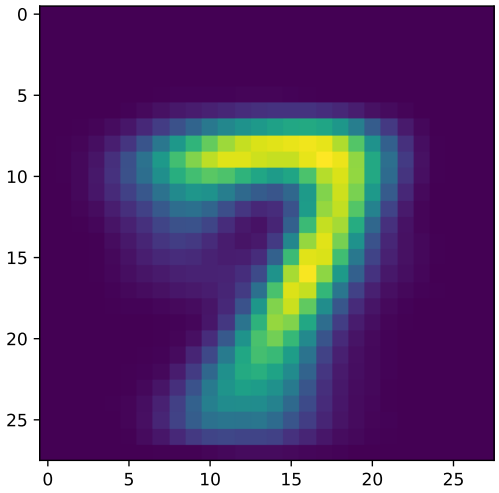


TABLE DES MATIÈRES

1 K-means demo 2
2 Number of clusters for K-means 2
2.1 Visual method . . . 2
2.2 Algorithmic heuristics . . . 2
2.3 Changing the data . . . 3
3 Application of vector quantization to classification 3
4 Compatibility graphs and custom metrics 9
5 Influence on data scaling on the convergence of SGD 13

1 K-MEANS DEMO

In **k_means_demo/**, you can find an example of k-means algorithm, implemented with numpy, with 3 clusters in dimension 2. If the initialization is kept random, the results will be random but will typically look like figure 1.

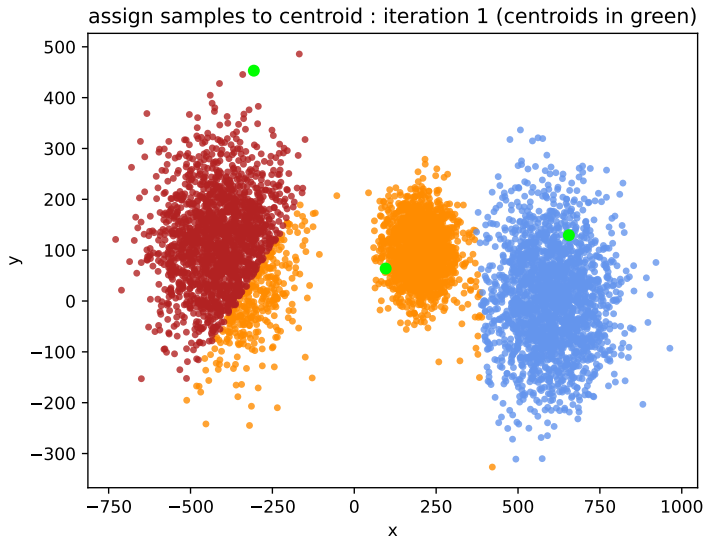


FIGURE 1 – An iteration step in the k-means algorithm

2 NUMBER OF CLUSTERS FOR K-MEANS

In this exercice we study heuristics to find relevant number of clusters for a k-means algorithm.

— folder : **exercice_1_k_means/**

— template file : **k_means_heuristics.py**

A company has gathered data about its customers and would like to identify similar clients, in order to propose relevant products to new clients, based on their features. This can be represented as a clustering problem. The data are stored in **data.npy**. They are 4 dimensional. We would like to study methods to automatically find a relevant number of clusters in this dataset for the K-means algorithm.

2.1 Visual method

Is there a direct, visual way to have an idea of a relevant number of clusters for these 4-dimensional data ?

2.2 Algorithmic heuristics

We would like to see if algorithmic heuristics are consistent with the previous results.

Experiment with the following heuristics, metrics and tools in order to obtain a suggested number of clusters :

— possible metrics to monitor :

  — inertia :

    https://scikit-learn.org/stable/modules/clustering.html#k-means

  - silhouette score
    https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient
  - Calinksi-Harabasz score :
    https://scikit-learn.org/stable/modules/clustering.html#calinski-harabasz-index
- Knee detection tools :
  - https://github.com/arvkevi/kneed
  - https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

2.3 Changing the data

You can generate the data again in order to study the behavior of the previous heuristics on a dataset with different statistical properties.

3 APPLICATION OF VECTOR QUANTIZATION TO CLASSIFICATION

In this section, we build a classifier based on an unsupervised preprocessing of the data. We will apply **vector quantization** to the MNIST classification problem. Namely, we will compute **an average representer** (prototype) for of each class, and for a new sample, predict the class of the nearest prototype. Some of the prototypes are represented in figures 2, 3, 5, 4, 7, 6. We might not expect a very high classification accuracy with this classifier, but this is rather interesting as a simple benchmark for this classification problem, on which we can achieve more than 0.98 accuracy with neural networks for instance, in a couple of minutes. We can see some examples of misclassified digits in figure 11 and 12.

https://en.wikipedia.org/wiki/Vector_quantization
You can fetch the data using vector_quantization/fetch_data.py.

A **Gaussian blur** of the input digits (see figures 8, 9, 10) might slightly improve the classification performance.

Implement this classification method based on vector quantization in order to classify the MNIST dataset and evaluate the test accuracy of this method. You might optionnaly use a hyperparameter optimization method in order to find a good value of the σ parameter of the GaussianBlur method of OpenCV.
https://docs.opencv.org/4.x/d4/d86/group__imgproc__filter.html#gaabe8c836e97159a9193fb0b11a

For this exercise, a decomposition of the dataset into a train/validation/test might be enough, but you can also use a cross validation.

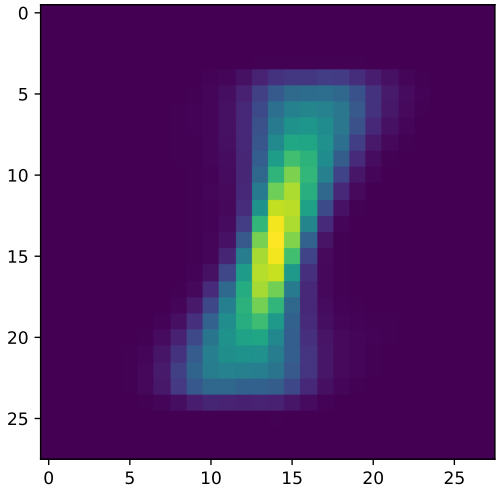


FIGURE 2 – Average of the 1 class (prototype).

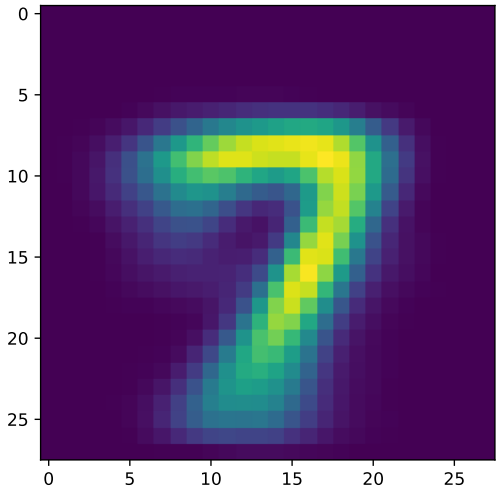


FIGURE 3 – Average of the 7 class (prototype).

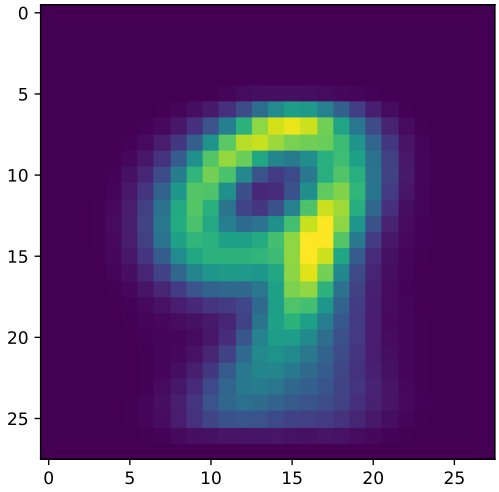


FIGURE 4 – Average of the 9 class (prototype).

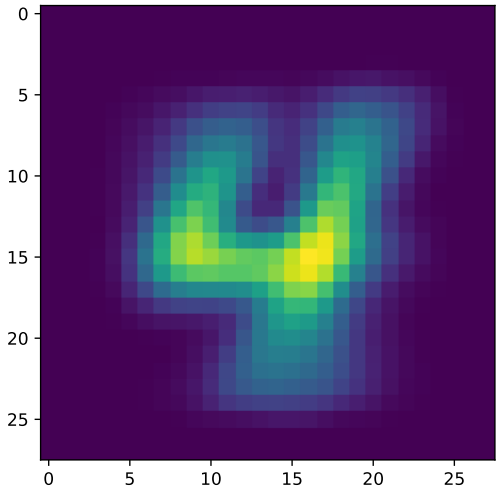


FIGURE 5 – Average of the 4 class (prototype).

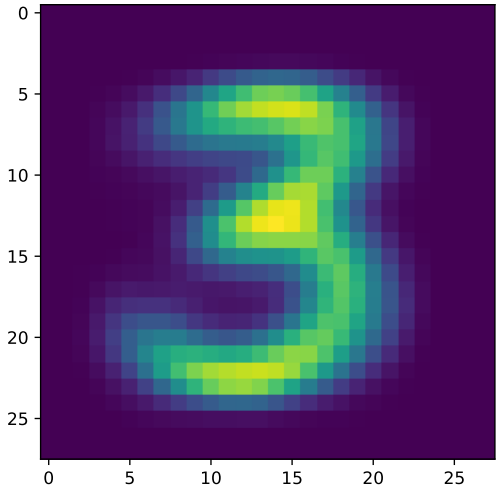


FIGURE 6 – Average of the 3 class (prototype).

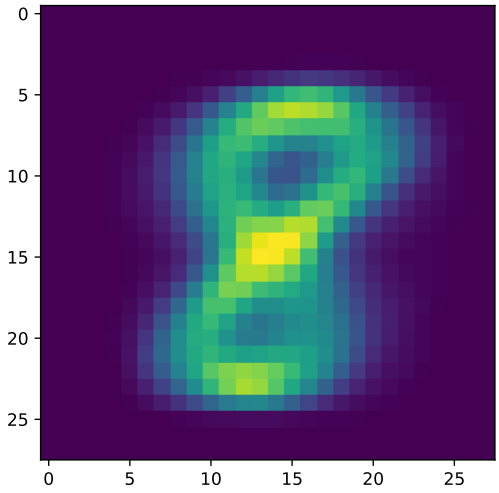


FIGURE 7 – Average of the 8 class (prototype).

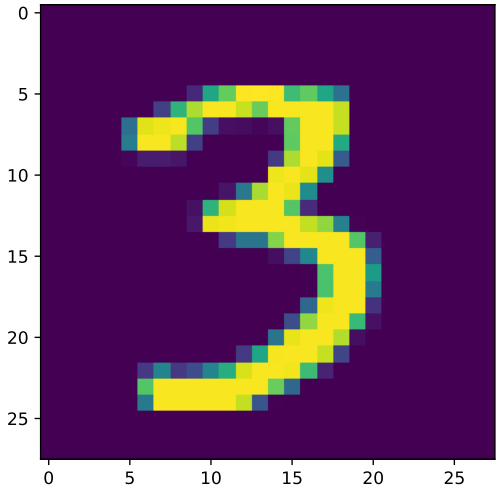


FIGURE 8 – A digit from the dataset.

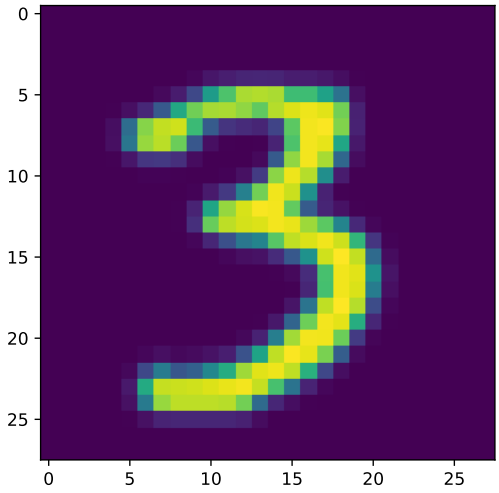


FIGURE 9 – The same digit, blurred with $\sigma = 0.5$ (Gaussian blur)

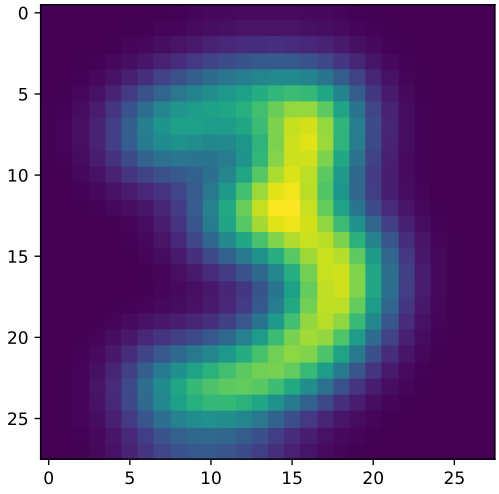


FIGURE 10 – The same digit, blurred with $\sigma = 2$

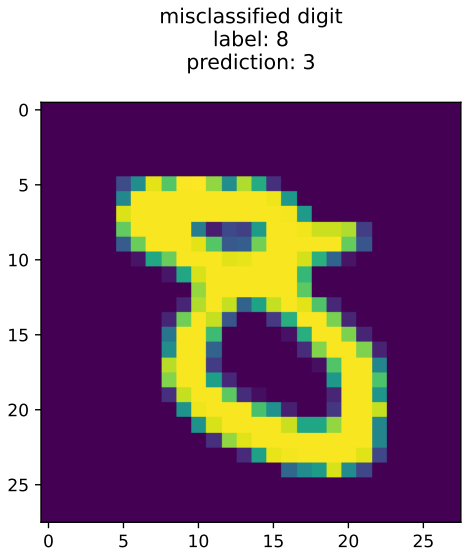


FIGURE 11 – Example of misclassified digit

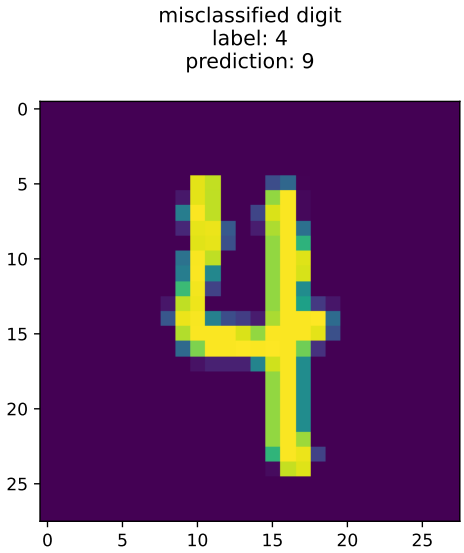


FIGURE 12 – Example of misclassified digit

4 COMPATIBILITY GRAPHS AND CUSTOM METRICS

In this exercise we build a custom distance in a dataset in order to observe a given compabtility graph (a graph where there is an edge between nodes that are closer than some threshold value). This illustrates the fact that the choice of the metric determines which samples are considered as similar or not in a dataset, which can be very important for practical applications, especially in unsupervised learning.

Even for geometric (and thus numerical data), the classical euclidean distance is not the only available metric. If we take a look at the documention of **cdist** from scipy or **numpy.linalg.norm**, we see that many metrics exist.

`https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.cdist.html`

`https://numpy.org/doc/stable/reference/generated/numpy.linalg.norm.html`

Use the notebook **build _graphs_geometric_data.ipynb**. in order to build compabtility graphs for the data contained in **data/data.npy** (displayed in figure 13), in order to obtaine the graphs shown in figures 14, 15, 16, 17, 18. You will need to choose the right **metric** for each graph. Try to think of the metric only mentally **before** implementing it !

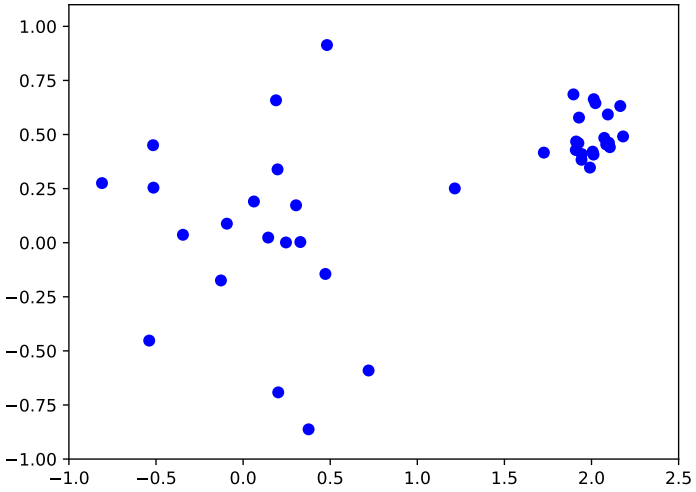


FIGURE 13 – The data to build compabtility graphs from.

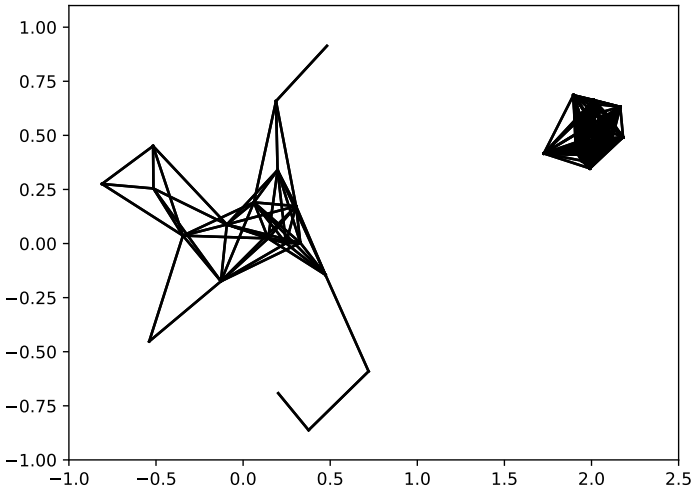


FIGURE 14 – Graph 1

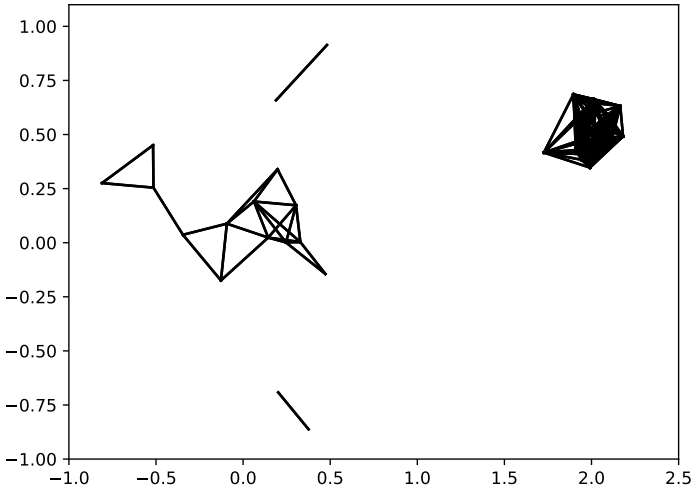


Figure 15 – Graph 2

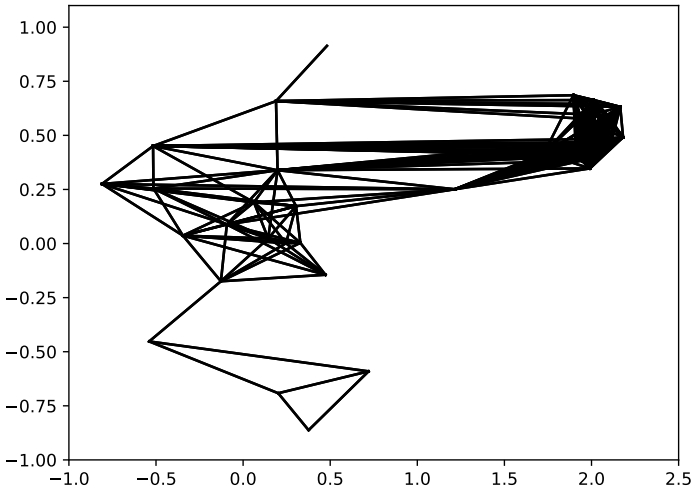


Figure 16 – Graph 3

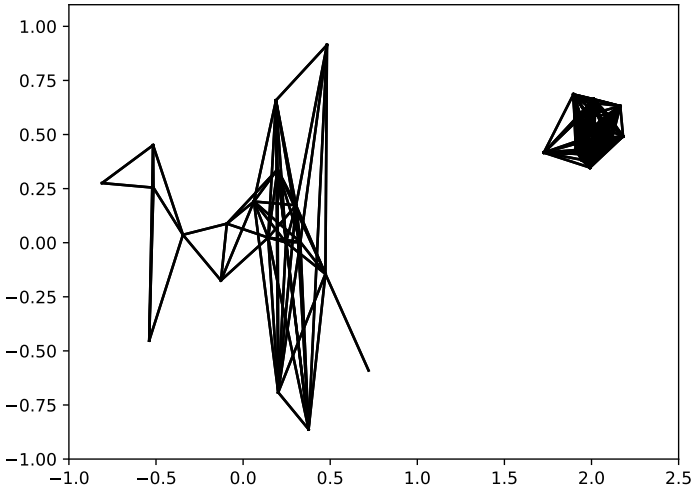


Figure 17 – Graph 4

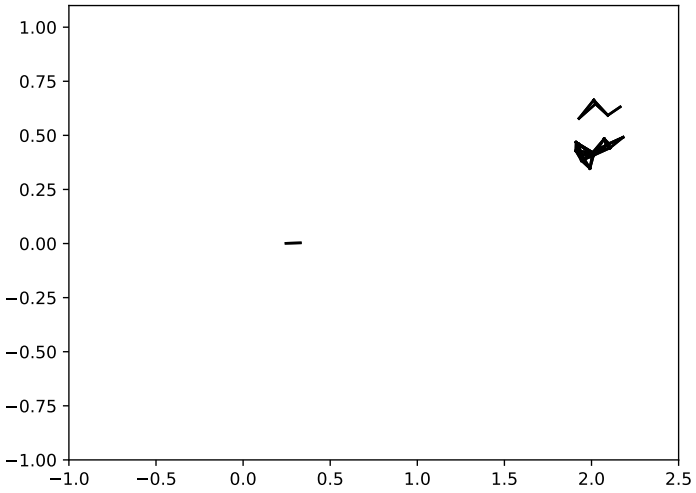


Figure 18 – Graph 5

5 INFLUENCE ON DATA SCALING ON THE CONVERGENCE OF SGD

We would like to perform a binary clasification on the following 3 datasets (figures 19, 20, 21). Each one has a different difficulty for a linear classifier, such as a SVM. We also note that the scales are different on the x and y axes.

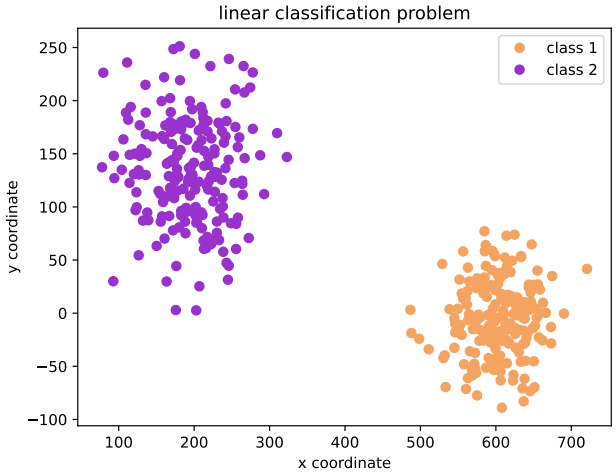


FIGURE 19 – Dataset 1

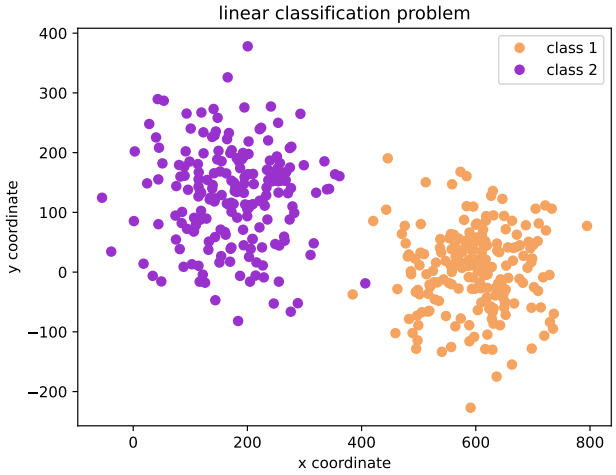


FIGURE 20 – Dataset 2

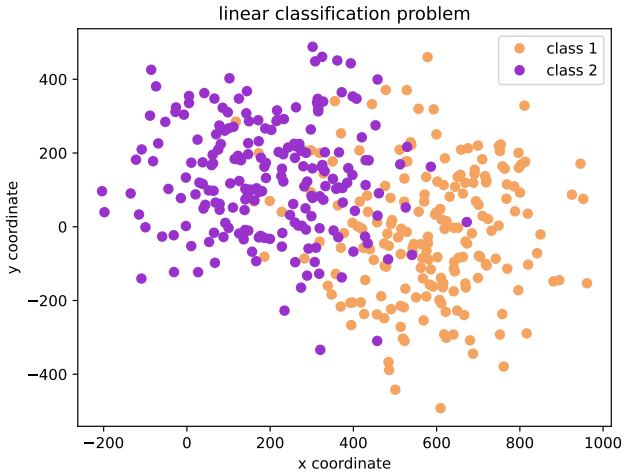


FIGURE 21 – Dataset 3

The data are located in **svm_sgd/data/**.

Data standardization consists in transforming the data so that each feature (each column) is centered (zero mean) and has a variance equal to 1. It is experimentally often noticed that algorithms trained by SGD give a better performance (generalization error) when the data are standardized.

`https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-scaler`.

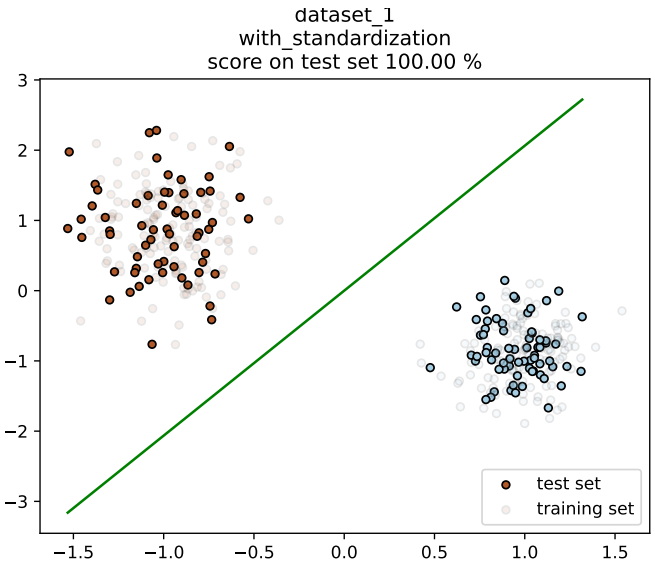
Perform a linear classification on these data, optimized by SGD, and compare quality of the results with and without preprocessing the data by standardizing them. You may use scikit-lean to do it, in particular :

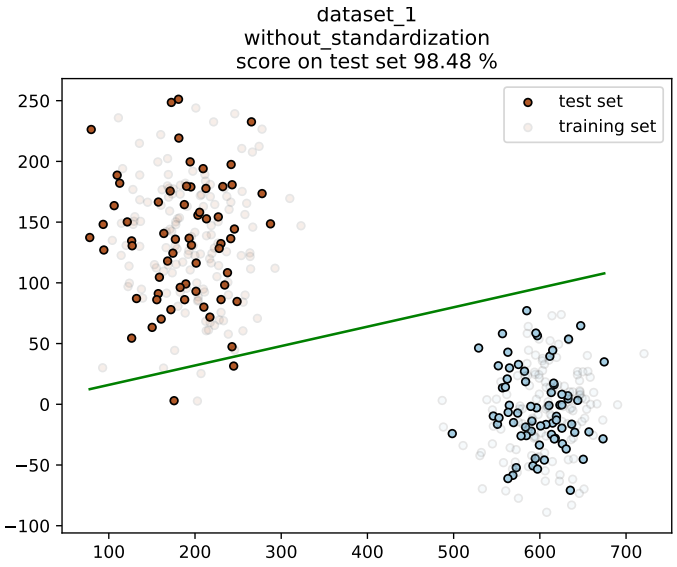
— `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html`. By default, this class trains a liner SVM (no kernel).

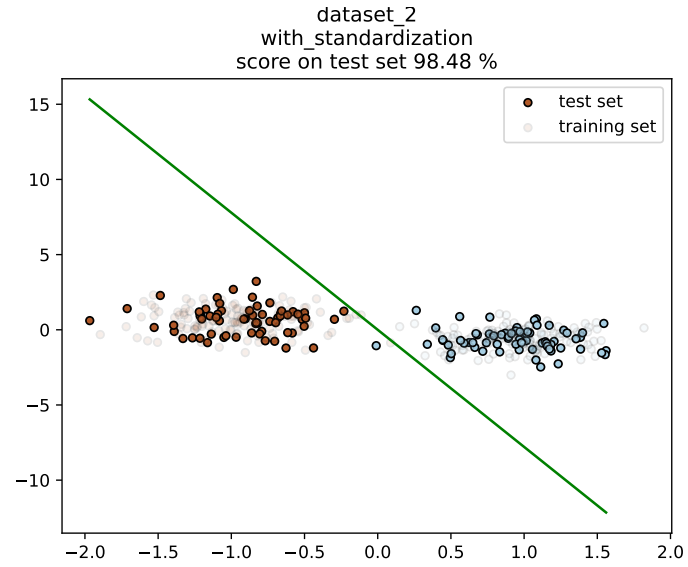
— `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

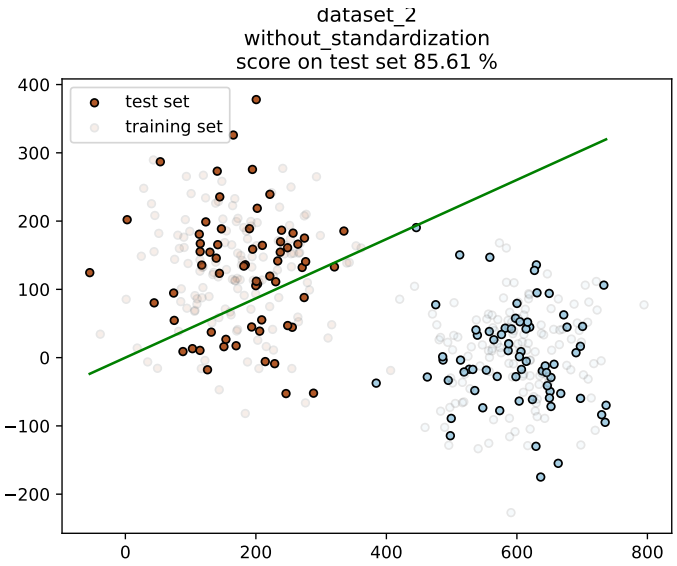
— `https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html`

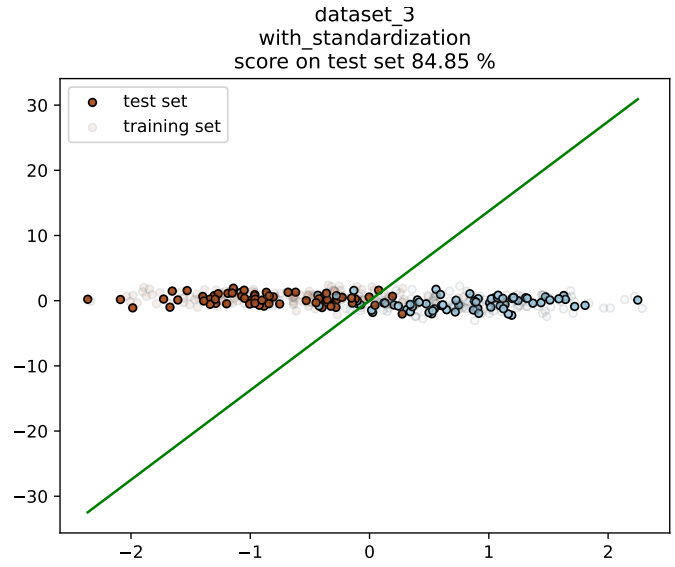
You should obtain results like the following figures.

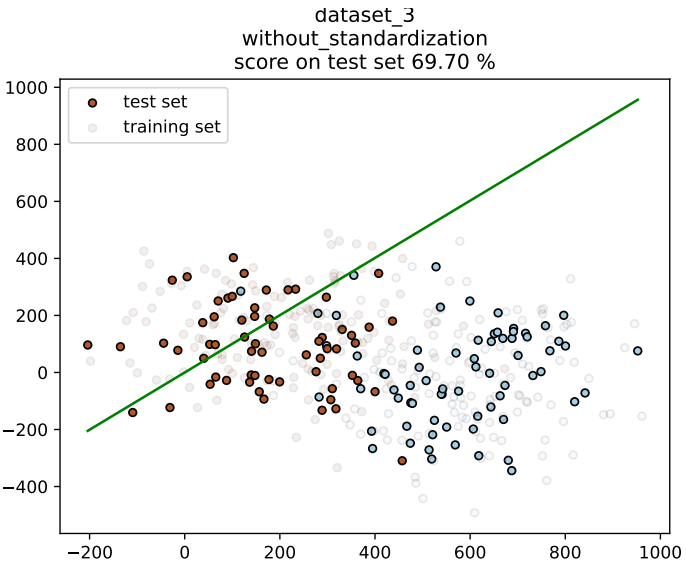



dataset_1
without_standardization
score on test set 98.48 %
test set
training set
250
200
150
100
50
0
−50
−100
100
200
300
400
500
600
700


dataset_2
with_standardization
score on test set 98.48 %
test set
training set
15
10
5
0
−5
−10
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0


dataset_2
without_standardization
score on test set 85.61 %
test set
training set
400
300
200
100
0
−100
−200
0
200
400
600
800





In this example, we saw that data standardization may give an improved performance, on a linear SVM trained by SGD. You can perform the same study on different datasets such as the toy datasets from scikit or other datasets.